Machine Learning 6.867 - Project

December 10, 2015

1 Introduction

We will first code this project in Julia. Paper [2] provides a walkthrough of MFVB applied to logistic regression, and sample MATLAB code is available on Github. Paper [1] includes a description of computational experiments with LRVB applied to a Normal-Poisson model, a linear random effects model, a multivariate normal mixture model, and the MNIST classification problem using a clustering approach. In addition, we have access to a large subset of classification problems from the UCI Machine Learning Repository (75 total), which have been pre-processed and do not require data cleaning. For this project, we will focus on the mathematical derivation of LRVB for logistic regression and its implementation on small-scale problems and simulated datasets. In addition, if we observe significant differences using the out-of-sample accuracy measure on these small-scale problems, then we can easily scale to run our methods on these 75 datasets.

A recent paper by Giordano, Broderick and Jordan (2015) proposes a linear correction term derived analytically to improve the covariance estimate for MFVB. The new proposed method is called Linear Response Variational Bayes (LRVB). Here, we consider applying this new method on logistic regression and compare its performance to the benchmark of MFVB and regular logistic regression.

2 Mean-Field Variational Bayes

Mean-field variational Bayes (MFVB) is a method for approximating the posterior distribution. In general, we have unknown parameters $w_1, w_2, \ldots, w_n$ that we have priors on, and our objective is to find the joint distribution $p(w_1, w_2, \ldots, w_n)$. Assuming that our approximate distribution is in the family $Q = \{q : q(w_1, w_2, \ldots, w_n) = q(w_1)q(w_2)\ldots q(w_n)\}$, we find $q^* \in Q$ that minimizes the KL-divergence with p, i.e. $q^* = \min KL(q||p)$. In particular, for logistic regression, the analytical form of the posterior is unknown and has been approximated with MFVB in the literature [2]. We use local variational bounds on the conditional probability using the convexity of the logarithm function. In particular, we use a variational treatment based on the approach of Jaakkola and Jordan (2000). This approach consists of approximation the likelihood function of the logistic regression, governed by the sigmoid function, by the exponential of the a quadratic form, leading to a gaussian approximation of the posterior distribution. More explicitly, if $y \in \{-1, 1\}$ is a target variable for a data vector x then the likelihood function of the target variable y is :

$$p(y|x, w) = \sigma(yw^T x) \tag{1}$$

with w being the logistic regression weight, and $\sigma(x) = \dfrac{1}{1 + \exp(-x)}$ the sigmoid function. Using a transformation of a the logarithm of the sigmoid and the concept of convex duality, we get :

$$\sigma(x) \geq \sigma(\xi) \exp((x - \xi)/2 - \lambda(\xi)(x^2 - \xi^2)) \tag{2}$$

where

$$\lambda(\xi) = \frac{1}{2\xi}[\sigma(\xi) - \frac{1}{2}] = \frac{1}{4\xi} \tanh(\frac{\xi}{2})$$

and ξ is a variational parameter.

Therefore, if we let $a = w^T x$ we get:

$$p(y|x, w) \geq e^{ya}\sigma(\xi)\exp\{-(\xi + a)/2 - \lambda(\xi)(a^2 - \xi^2)\} \tag{3}$$

To every training set observation (x_n, y_n), there is a variational parameter ξ_n associated. We apply the bound above to each of the terms in the likelihood function. Let $Y = [y_1, y_2, \ldots, y_n]^T$ and the X be the data matrix, then the likelihood function is:

$$p(Y|X, w) = \prod_{i=1}^{N} p(y_i|x_i, w) = \prod_{i=1}^{N} \sigma(yw^T x) \tag{4}$$

and thus we obtain the following bound on the joint distribution on y and w, assuming a prior $p(w)$ on w:

$$p(Y, w|X) = p(Y|X, w)p(w) \geq h(w, \xi)p(w) \tag{5}$$

and $h(w, \xi) = \prod_{i=1}^{N} e^{y_i w^T x_i}\sigma(\xi_i)\exp\{-(\xi_i + w^T x_i)/2 - \lambda(\xi_i)((w^T x_i)^2 - {\xi_i}^2)\}$

However, the variational Bayes approximation is known to underestimate the covariance matrix of the posterior distribution, and this estimate can be made arbitrarily bad for simulated examples with 2 or more dimensions [1].

3 Markov-Chain Monte Carlo

To evaluate the quality of the covariance estimates produced by our method, we used Markov-Chain Monte Carlo (MCMC) as a benchmark for the "true" distribution of the logistic regression weights (w_0, w_1, w_2). We used the `R` package `MCMCpack` with an improper uniform prior, 10,000 iterations, and burn-in rate of 1,000 iterations. We also compared to an MCMC simulation with a normal prior on the weights $\mathcal{N}(\mathbf{0}, 1000I)$ and found similar results. Figures **??** and **??** show the progression of the MCMC algorithm assuming each prior.

To visualize the joint distribution of the logistic regression weights, we plot the MCMC results for the values of w_1 and w_2. In addition, we

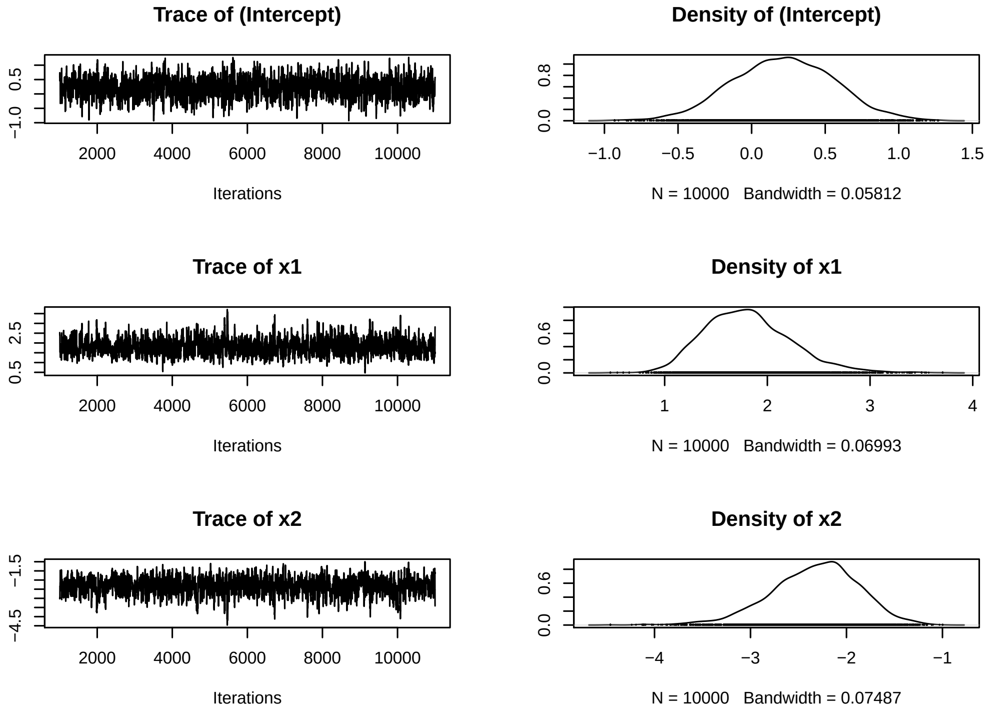

Figure 1: MCMC simulations of logistic regression weights for dataset 1, and corresponding marginal density plots, assuming an improper uniform prior. 10,000 iterations total.

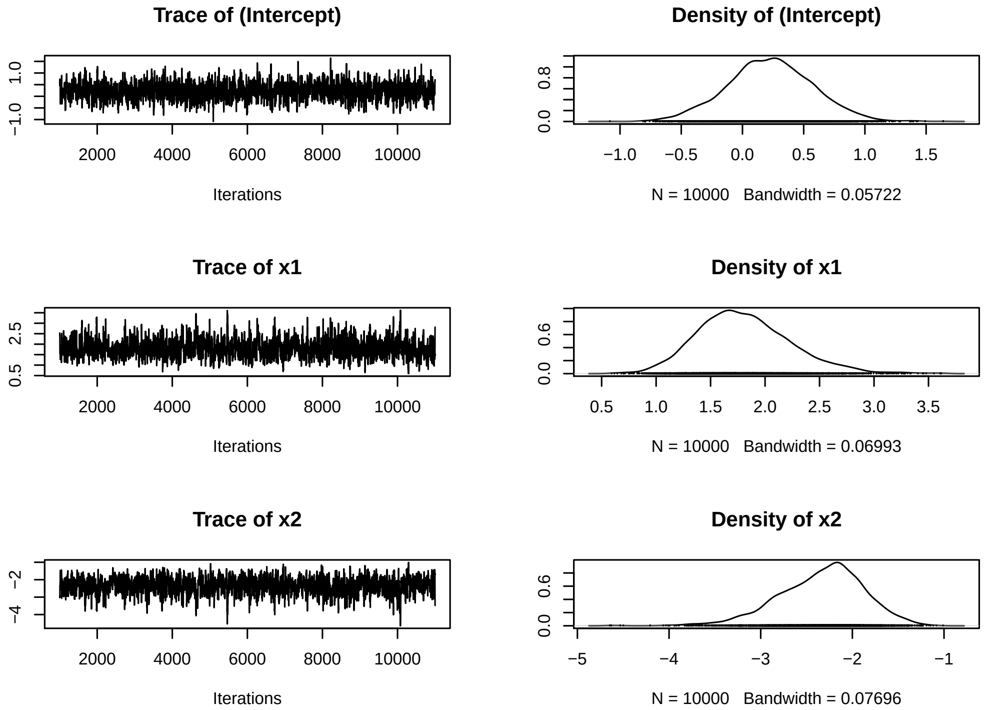

Figure 2: MCMC simulations of logistic regression weights for dataset 1, and corresponding marginal density plots, assuming a normal prior $\mathcal{N}(\mathbf{0}, 1000I)$. 10,000 iterations total.

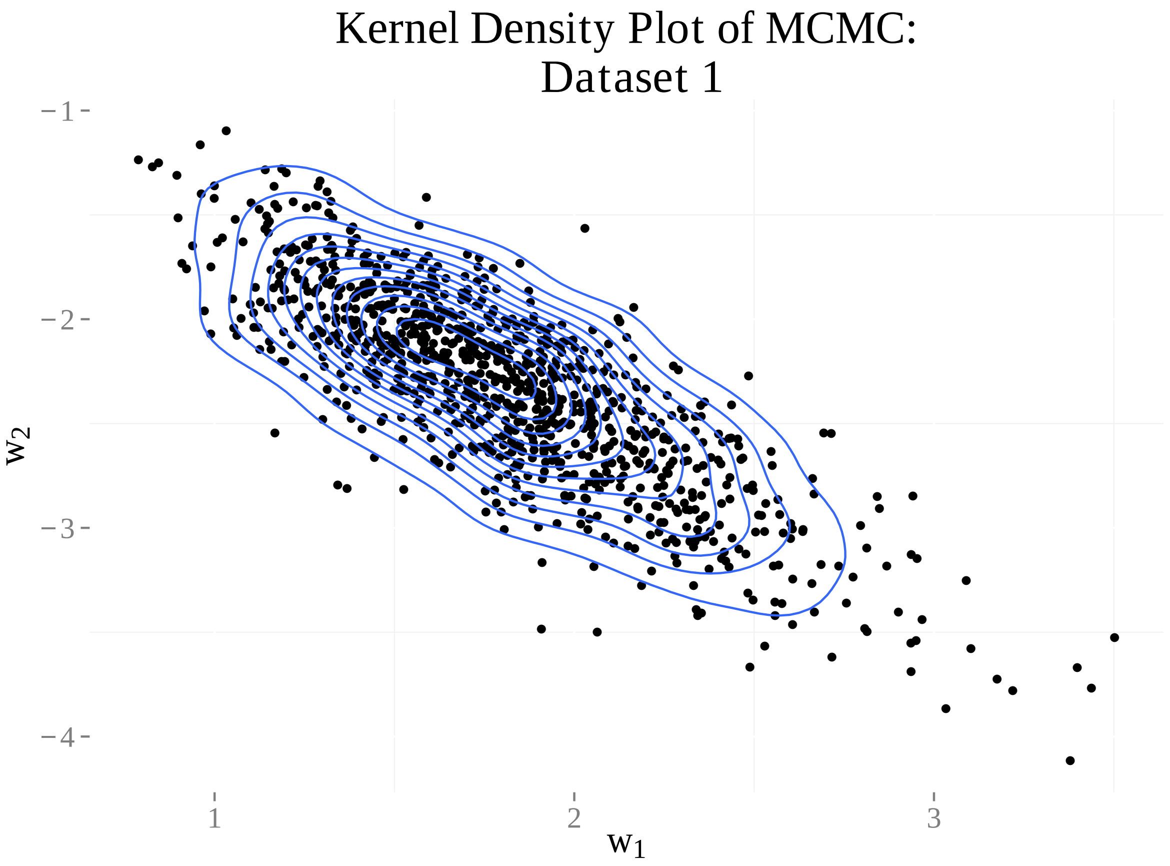

Figure 3: MCMC simulations of logistic regression weights for dataset 1, and corresponding kernel density plot, assuming an improper uniform prior. Subset of 1,000 out of 10,000 total iterations shown.

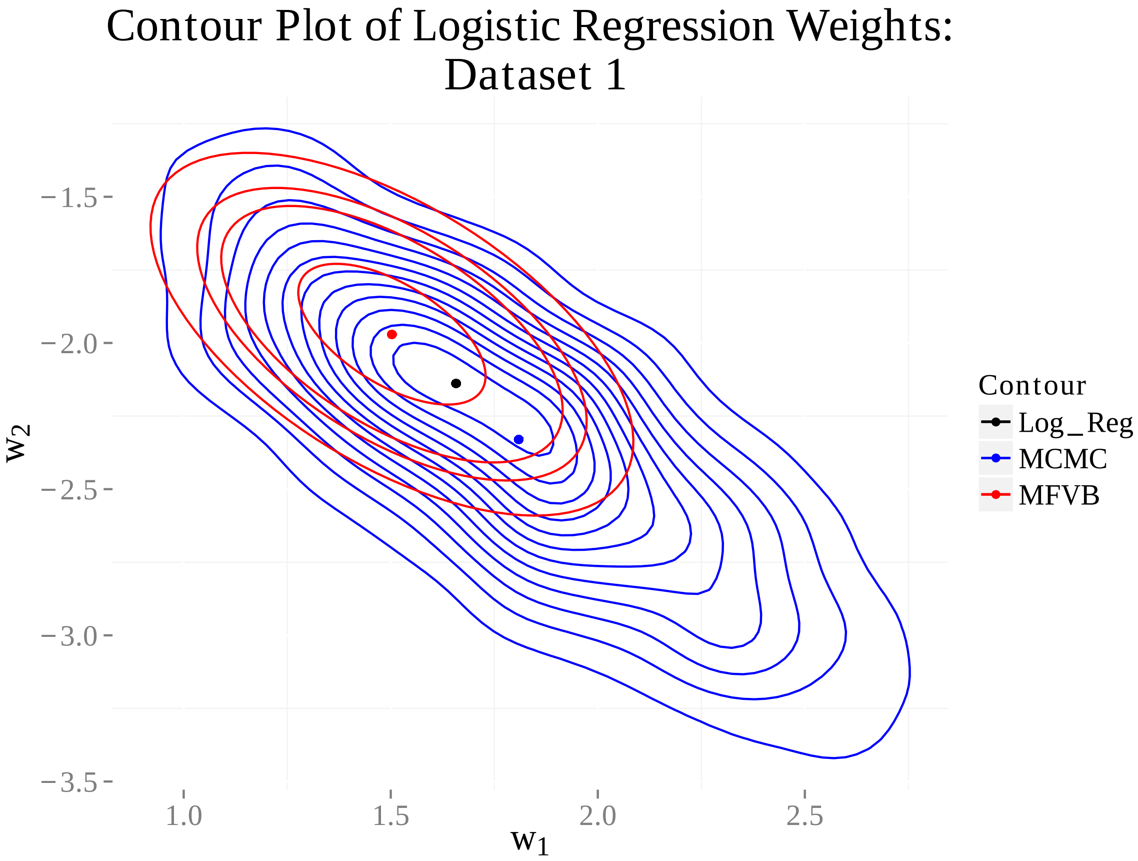

Figure 4: Comparison of logistic regression point estimate, kernel density plot of MCMC simulations, and posterior density of MFVB logistic regression (vectorized function), for dataset 1. Contours of MFVB logistic regression indicate 50%, 90%, 95%, and 99% confidence intervals for the bivariate normal $\mathcal{N}(\mathbf{w}_N, \mathbf{V}_N)$. Mean values for MCMC and MFVB are also included.